Research Results based on the CLARIN Infrastructure

Bente Maegaard
University of Copenhagen
bmaegaard@hum.ku.dk
This presentation builds on joint work in the CLARIN-DK team, as well as work done by researchers.

The CLARIN-DK team:

- Lene Offersgaard
- Dorte Haltrup Hansen
- Lina Henriksen
- Bart Jongejan
- Costanza Navarretta
- Sussi Olsen
- Claus Povlsen
- Mitch Seaton
- And earlier: Hanne Fersøe
CLARIN in 6 bullets

• **CLARIN** is the Common Language Resources and Technology Infrastructure

• **ESFRI** ERIC status since 2012, Landmark since 2016

• CLARIN provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond

• to **digital language data** (in written, spoken, video or multimodal form)

• and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located

• through a **single sign-on** environment.
• CLARIN Vision
  All digital language resources and tools from all over Europe and beyond are accessible through a single sign-on environment for the support of researchers in the humanities and social sciences.

• CLARIN Mission
  Create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences.
CLARIN ERIC: members and centres

A consortium of:

- 19 members: AT, BG, CZ, DE, DK, DLU, EE, FI, GR, HU, IT, LT, LV, NL, NO, PL, PT, SE, SI

- 2 observers: FR, UK;

- >40 centres
Sharing data and tools

• The driving principle for CLARIN – and for any infrastructure – is sharing
• CLARIN is for sharing data and tools, and also knowledge
• This way CLARIN also contributes to the European Open Science Policies
• CLARIN is very close to adhering to the FAIR principles.
The research infrastructure is the result of an integrated collaboration between the central level and the national level.

All data deposited by the Danish team is available to the rest of CLARIN, through the Virtual Language Observatory (VLO).

Similarly, we can find data from the other countries.

The same is true for tools and services.

Normally, if you are on the Danish website then you will choose the tools you want to use – see the demo.

Additionally, a specific platform has been built centrally which tells which tools are available for a selected data resource – very useful if you do not know what is available (Language Resource Switchboard).
The Danish contribution to VLO

CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the search bar below to start searching through hundreds of thousands of language resources, or continue to browse everything and use facets to narrow down to your area of interest or discover new resources.

See all records  Learn more

Search

Showing 1 to 10 of 158262 results within selection for  Danish ×

Results per page:  10

DK-CLARIN LSP Corpus - Health domain II

"DK-CLARIN LSP Corpus - Health domain II" is a part of the Danish DK-CLARIN LSP corpus consisting of seven sub-corpora from following subject domains: Agriculture, Construction, Economics, Environment, Health, IT and Nanotechnology. Texts in the Health and Medicine Domain come from sundhed.dk and have been...

DK-CLARIN LSP Corpus - Construction domain

"DK-CLARIN LSP Corpus - Construction domain" is a part of the Danish DK-CLARIN LSP corpus consisting of seven sub-corpora from following subject domains: Agriculture, Construction, Economics, Environment, Health, IT and Nanotechnology. Texts in the Construction Domain come from Statens Byggeforsknin...

DK-CLARIN LSP Corpus - Economics domain

...
Anne Sofie Jacobsen is working on a PhD on Academic language. Her work is corpus-based. She collected her corpus from academic articles from different domains (science, health, humanities etc.) and compared it to existing corpora on general language.
• **Purpose:** Empirically based description of academic vocabulary in Danish

• **Research questions:**
  - Which lexemes can be identified as part of Danish academic language?
  - What is the distribution, meaning, function, form and structure of these lexemes?
  - Are there differences in the distribution, meaning, function of these lexemes in different scientific disciplines?
Academic language and academic vocabulary

• "Academic language is the specialized language, both written and oral, of academic settings that facilitates communication and thinking about disciplinary content." (Nagy and Townsend, 2012:92)

• The vocabulary is often abstract, and is part of creating the academic discourse ("the glue of academic language use")
Use of CLARIN.DK

• What she needed:
  • Conversion from pdf to txt
  • Tokenisation
  • Lemmatisation
  • PoS tagging
  • File format for SketchEngines
  • Treating more files at a time
  • Keep the unique file name

• At the time these functions were not integrated the same easy way as now
Some of the most frequent lemmas

<table>
<thead>
<tr>
<th>English</th>
<th>Danish</th>
<th>English</th>
<th>Danish</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>central</strong></td>
<td><strong>central</strong></td>
<td><strong>fokus</strong></td>
<td><strong>focus</strong></td>
</tr>
<tr>
<td><strong>adj</strong></td>
<td><strong>(adj)</strong></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>fænomen</strong></td>
<td><strong>phenomenon</strong></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>perspektiv</strong></td>
<td><strong>perspective</strong></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>gennemsnit</strong></td>
<td><strong>mean</strong></td>
<td><strong>henholdsvis</strong></td>
<td><strong>respectively</strong></td>
</tr>
<tr>
<td><strong>(n)</strong></td>
<td></td>
<td><strong>basere</strong></td>
<td><strong>base</strong></td>
</tr>
<tr>
<td><strong>(adv)</strong></td>
<td></td>
<td><strong>(v)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>praksis</strong></td>
<td><strong>practice</strong></td>
<td><strong>begreb</strong></td>
<td><strong>concept</strong></td>
</tr>
<tr>
<td><strong>(n)</strong></td>
<td></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>dialog</strong></td>
<td><strong>dialogue</strong></td>
<td><strong>interview</strong></td>
<td><strong>interview</strong></td>
</tr>
<tr>
<td><strong>(n)</strong></td>
<td></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>resultat</strong></td>
<td><strong>result</strong></td>
<td><strong>undersøge</strong></td>
<td><strong>investigate</strong></td>
</tr>
<tr>
<td><strong>(n)</strong></td>
<td></td>
<td><strong>(v)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>ligeledes</strong></td>
<td><strong>likewise</strong></td>
<td><strong>analyse</strong></td>
<td><strong>analysis</strong></td>
</tr>
<tr>
<td><strong>(adv)</strong></td>
<td></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Eksempelvis</strong></td>
<td><strong>as an example</strong></td>
<td><strong>element</strong></td>
<td><strong>element</strong></td>
</tr>
<tr>
<td><strong>(adv)</strong></td>
<td></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>vurdere</strong></td>
<td><strong>assess</strong></td>
<td><strong>(v)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>felt</strong></td>
<td><strong>field</strong></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>forudsætning</strong></td>
<td><strong>assumption</strong></td>
<td><strong>effekt</strong></td>
<td><strong>effect</strong></td>
</tr>
<tr>
<td><strong>(n)</strong></td>
<td></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>relevant</strong></td>
<td><strong>relevant</strong></td>
<td><strong>(adj)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>tendens</strong></td>
<td><strong>tendency</strong></td>
<td><strong>(n)</strong></td>
<td></td>
</tr>
</tbody>
</table>
Results

• The research question had nothing to do with language technology
• But language technology and the CLARIN technology and expertise supported the corpus work.
Corpus search in folk songs

- (Hanne Ruus and Dorthe Duncker, NorS, KU)
- Folk songs 1553-1700
- Two versions: original spelling and transcription to modern Danish, see more at: duds.nordisk.ku.dk

Den ældste danske viseoverlevering
Håndskrift: Hjertebogen
Nummer: 68
Side: 143 v
Titel: Nilus og Adelus
Trykt: T
Udgivet i: DgF: 481 Aa VIII 307-308
Kommentar: VIII 368a
Generelle oplysninger:

_Efter Teksten tilføjor Skriveren:_ O Gud hussuale alle bedrouet hierter och unne dem alle thett, Gud er tæckeligt,
onch dem er nøtteligt, bode tiill liiff och siell. Amen.

1
Der star yett tre y kongens gardt,
der står et træ i kongens gård,
er woxsen mett rosser och lillier;
er voksen med roser og liljer;
hér Nielaus och stalten Adelutz
hér Nielaus og stolten Adelus
thy hade tiill-samen god wyle.
de havde tilsammen god vilje.
:: Sorren ber thy for elskoff. ::
:: sorgen bør de for elskov. ::
A search for the word "lykke" in a Danish corpus containing 65 matches. The top matches include works by Jens Billes håndskrift (Jens Bille), 1555-1559, and Anna Munks håndskrift.

The frequencies of the word in various works are shown: Dronning Sophias visebog (15.4%), Hjertebogen (18.5%), Jens Billes håndskrift (13.8%), Karen Brahes folio, ældre del (4.6%), Langebeks kvart (21.5%), Rentzells håndskrift (3.1%), Svanings håndskrift I (10.8%), and Svanings håndskrift II (3.1%).
• Gesta Danorum (GD) is written by Saxo around 1200
  – a compilation - based on oral tradition - of mythical stories about Danish legendary kings
  – and a historical monograph describing contemporary political events.

• GD consists of 16 books.
  – The first section consists of the books dealing with Norse mythology and the second section describes the political events, and the introduction of Christianity in Denmark.

• Where is the borderline?
  – Traditionally GD has been split between book 9 and 10: books 1-9 describe the heathen era, and books 10-16 the transition to Christianity
  – But in 1969, a competing hypothesis was launched: books 1-8, and books 9-16. This was based on a comparison with other documents from the same time etc. Not immediately accepted.
• A translation from Latin to Danish (Friis-Jensen and Zeeberg, 2005) was used.

• Some elements characterizing Christian language were identified, e.g.
  – “helgen” (en: saint),
  – ”kristen” (en: Christian), ”kristendom” (en: Christianity)
  – ”synd” (en: sin) but importantly not ”synd” as an adjective like in ”det var synd for hende” (en: it was a pity for her).

• The IMS Open Corpus Workbench (CWB) was used.
Result of the search for Christian words

276 matches. (Click on blue items to display examples.)

- Saxos Danmarkshistorie. (a) Første bog: 2.9% f=8
- Saxos Danmarkshistorie. (b) Anden bog: 0.7% f=2
- Saxos Danmarkshistorie. (c) Tredje bog: 2.2% f=6
- Saxos Danmarkshistorie. (d) Fjerde bog: 0.7% f=2
- Saxos Danmarkshistorie. (e) Femte bog: 1.1% f=3
- Saxos Danmarkshistorie. (f) Sjette bog: 0.7% f=2
- Saxos Danmarkshistorie. (g) Syvende bog: 0.7% f=2
- Saxos Danmarkshistorie. (h) Ottende bog: 1.8% f=5
- Saxos Danmarkshistorie. (i) Niende bog: 8.0% f=22
- Saxos Danmarkshistorie. (j) Tiende bog: 14.5% f=40
- Saxos Danmarkshistorie. (k) Elleve bog: 13.8% f=38
- Saxos Danmarkshistorie. (l) Tolvte bog: 6.9% f=19
- Saxos Danmarkshistorie. (m) Trettende bog: 6.2% f=17
- Saxos Danmarkshistorie. (n) Fjortende bog: 31.9% f=88
- Saxos Danmarkshistorie. (o) Femtende bog: 5.8% f=16
- Saxos Danmarkshistorie. (p) Sekstende bog: 2.2% f=6
• The occurrence figures cannot be taken directly, as the books are of different length. Book 10 is twice as long as book 9, so there is no big difference in the relative frequency

• Book 8 is roughly the same size as book 10, so the relative frequency of the Christian words is much less in book 8

• So, the split is between book 8 and 9 as Skovgaard-Petersen (1969) suggested.
Widespread use of the lemmatiser and PoS tagger

• Apart from these Danish uses of the infrastructure and the tools, we should mention that the lemmatiser and tagger have been used by many teams in the world. Just to mention two:
• Tim Tangherlini, UCLA, used the lemmatiser for his works on Danish
• The company behind SketchEngine used these tools as well. Quote from an email from the director: “Vit has now processed 2 billion words of Danish with your tools.” This is an example of a commercial use.
Distant reading of Danish literature
Tools chains (see demo)

Automatically annotate uploaded files with linguistic information.

Choose one or more files to convert to TEIP5-DKCLARIN. All files must have the same file format (pdf, rtf, doc, docx, odt, txt or html).

Sætningssegmentering (u. tokeniserings)  Dansk  Engelsk
Sætningssegmentering (m. tokeniserings)  Dansk  Engelsk
Brill POS-tagger (inkl. tokeniserings og sætningssegmentering)  Dansk  Engelsk
Lemmatiser (i løbende tekst)  Dansk  Engelsk
Lemmatiser (i løbende tekst, inkl. POS-tags)  Dansk  Engelsk
Frekvensliste (lemma)  Dansk  Engelsk
Frekvensliste (ordformer)  Dansk  Engelsk
Frekvensliste (bigram, lemma)  Dansk  Engelsk
Frekvensliste (bigram, ordformer)  Dansk  Engelsk
Navnegenkender  Dansk
Navnegenkender (m. POS-tags)  Dansk
Corpus tool (see demo)
Final words

- CLARIN has two important parts: the international and the national. We have seen the importance and the contribution of both parts, as well as the integration.
- Several examples of specific research: two very different language studies, one historical study and finally an example which shows the high interest in using the tools worldwide.
- Finally, a few general descriptions with appetizers for the poster and demo session.